A Sparse PLS for Variable Selection when Integrating Omics data

Kim-Anh Lê Cao, Debra Rossow, Christèle C. Robert-Granié, Philippe Besse

A Sparse PLS for Variable Selection when Integrating Omics data

Kim-Anh Lê Cao[1,2], Debra Rossouw[3], Christèle Robert-Granié[2] and Philippe Besse[1]

Abstract

Recent biotechnology advances allow for the collection of multiple types of omics data sets, such as transcriptomic, proteomic or metabolomic data to be integrated. The problem of feature selection has been addressed several times in the context of classification, but has to be handled in a specific manner when integrating data. In this study, we focus on the integration of two-block data sets that are measured on the same samples. Our goal is to combine integration and simultaneous variable selection on the two data sets in a one-step procedure using a PLS variant to facilitate the biologists interpretation. A novel computational methodology called "sparse PLS" is introduced for a predictive purpose analysis to deal with these newly arisen problems. The sparsity of our approach is obtained by soft-thresholding penalization of the loading vectors during the SVD decomposition.

Sparse PLS is shown to be effective and biologically meaningful. Comparisons with classical PLS are performed on simulated and real data sets and a thorough biological interpretation of the results obtained on one data set is provided. We show that sparse PLS provides a valuable variable selection tool for high dimensional data sets.

Introduction

Motivation. Recent advances in technology enable the monitoring of an unlimited quantity of data from various sources. These data are gathered from different analytical platforms and allow their integration among different types, such as transcriptomic,

[1]Institut de Mathématiques, Université de Toulouse et CNRS (UMR 5219), F-31062 Toulouse, France

[2]Station d'Amélioration Génétique des Animaux UR 631, Institut National de la Recherche Agronomique, F-31326 Castanet, France

[3]Institute for Wine Biotechnology, University of Stellenbosch, Stellenbosch, South Africa

proteomic or metabolomic data. This integrative biology approach enables to understand better some underlying biological mechanisms and interaction between functional levels, if one succeeds in incorporating the several omics types of data, that are characterized by many variables but not necessarily many samples or observations. In this highly dimensional setting, the selection of genes, proteins or metabolites is absolutely crucial to overcome computational limits (from a mathematical and statistical point of view) and to facilitate the biological interpretation. Hence our quest of sparsity is motivated by the biologists needs, who want to separate the useful information related to the study from the non useful information, due to experiment inaccuracies. The resulting variable selection might also enable a feasible biological validation with a reduced experimental cost. We especially focus on the integration context, which is the main goal of omics data. For example, one biological study might aim at explaining the q metabolites by the p transcripts, that are measured on the same n samples. In this typical case, $n << p + q$.

In this study, we propose a sparse version of the PLS, that aims at combining selection *and* modelling in a one-step procedure for such problems. Our sparse PLS is based on soft-thresholding penalization and is obtained by penalizing a sparse SVD (Shen and Huang, 2007), using a hybrid PLS with SVD decomposition (Lorber et al., 1987). This approach deals with integration problems, that cannot be solved with usual feature selection approaches proposed in classification or discrimination studies where there is only one data set to analyse. Hence, multiple testing that looks for differentially expressed genes does not apply here, as well as other classification methods that were applied to transcriptomic data sets. In this latter case, many authors (among them: Guyon et al. 2002; Lê Cao et al. 2007) have applied feature selection methods to microarray data and have been proved to bring biologically meaningful genes lists. However, in our context, the feature selection aim has to be integrated with modelling, and very few approaches have been proposed to deal with these newly arisen problems, especially in a one-step procedure. In a two-block data sets setup, our aim is to *predict* one group of variables from the other group. Several approaches that seek linear combinations of both groups of variables can answer this biological problem. However, they are often limited by collinearity or ill posed problems, that require regularization techniques, such as l_1 (soft-thresholding, Lasso) or l_2 (Ridge) penalizations.

Background and related work. Partial Least Squares regression (PLS, Wold 1966) is a well known regression technique, mostly applied in chemometrics. Its stability property faced to collinear matrices gives PLS a clear superiority to CCA, multiple linear regression, ridge regression or other regression techniques. Furthermore, since Wold original approach, many variants have arisen (SIMPLS, de Jong 1993, PLS1 and 2, PLS-A, PLS-SVD, see Wegelin (2000) for a survey) that provide the user a solution for almost any problem. We will describe and discuss some of these variants in this study.

PLS has been successfully applied to biological data, such as gene expression (Datta, 2001), integration of gene expression and clinical data (with bridge PLS, Gidskehaug et al. 2007), integration of gene expression and ChIP connectivity data (Boulesteix and

Strimmer, 2005) and more recently for reconstructing interaction networks from microarray data (Pihur et al., 2008). We can also mention the study of (Culhane et al., 2003) who applied Co-Inertia Analysis (CIA, Doledec and Chessel 1994) from which PLS is a particular case, in a cross platform comparison in microarray data.

In the context of feature selection from both data sets, one closely related work proved to bring biologically meaningful results is the O2PLS model (Trygg and Wold, 2003), associated to variable selection in Bylesjö et al. (2007) for combining and selecting transcript and metabolite data in *Arabidopsis Thaliana* in a regression framework. O2PLS decomposes each data set in three structures (predictive, unique and residual). The most dominating correlation and covariance in both sample directions and variable directions is extracted and can be interpreted. Variable selection is then performed on the correlation loadings with a permutation strategy, hence with a two-step procedure.

More recently, Waaijenborg et al. (2008) and Chun and Keles (2007) both adapted Elastic Net regularization (Zou and Hastie, 2005) in the PLS, either in a canonical framework, or in a regression framework, by directly penalizing the optimization problem. Both approaches seem promising, as Chun and Keles (2007) demonstrated that the PLS consistency property does not hold when $n << p + q$. However, it would be useful to show the biological relevancy of their results. Nevertheless, their studies show the need of developing such integrative methods for biological problems.

Our contribution and results. We propose a sparse PLS approach that combines both integration and variable selection simultaneously on the two data sets, in a one-step strategy. We show that our approach is applicable on high-throughput data sets and bring more relevant results compared to PLS.

Outline of the paper. A brief introduction to PLS will be given, before describing the sparse PLS method. We detail how to add sparsity to PLS with a soft-thresholding penalization combined to SVD computation (Shen and Huang, 2007). We then assess the validity of the approach on one simulated and three real data sets, compare and discuss the results with a classical PLS approach. We also provide a full biological interpretation of the results obtained on a typical integrative study of wine yeast, that combines transcripts and metabolites. We show how sparse PLS highlights the most essential transcripts that are meaningfully related to the metabolites.

1 Methods

1.1 PLS

The PLS regression looks for a decomposition of centered (possibly standardized) data matrices X ($n \times p$) and Y ($n \times q$) in terms of components scores, also called latent variables: $(\xi_1, \xi_2 \dots \xi_H)$, $(\omega_1, \omega_2 \dots \omega_H)$, that are n-dimensional vectors, and associated

loadings: $(u_1, u_2 \dots u_H)$, $(v_1, v_2 \dots v_H)$, that are respectively p and q- dimensional vectors, to solve the following optimization problem (Burnham et al., 1996):

$$\max_{||u_h||=1,||v_h||=1} cov(X_{h-1}u_h, Yv_h) \tag{1}$$

where X_{h-1} is the residual (deflated) X matrix for each PLS component dimension $h = 1 \dots H$. Problem (1) is equivalent to solve: $\max cov(\xi_h, \omega_h)$.

Many PLS variants exist depending on the way X and Y are deflated, either in a symmetric ("PLS-mode A") or asymmetric way ("PLS2") (Tenenhaus, 1998; Wegelin, 2000), and the models will consequently differ. In this study, we will focus only on a regression framework.

In the case of a *regression mode* (asymmetric), the models of X- and Y-space are respectively (Hoskuldsson, 1988):

$$X = \Xi C^T + \varepsilon_1 \qquad Y = \Xi D^T + \varepsilon_2 = XB + \varepsilon_2 \tag{2}$$

where Ξ $(n \times H)$ is the matrix of PLS components ξ_h, B $(p \times H)$ is the matrix of regression coefficients. The column vectors of C and D are defined as $c_h = X_{h-1}^T \xi_h / (\xi_h' \xi_h)$ and $d_h = Y_{h-1}^T \xi_h / (\xi_h' \xi_h)$, and ε_1 $(n \times p)$ and ε_2 $(n \times q)$ are the residual matrices, $h = 1 \dots H$.

Another PLS alternatives exist depending if X and Y are deflated separately or directly using the cross product $M = X^T Y$ and the SVD decomposition. We will discuss these various approaches in sections 1.2 and 1.4. Note that in any case, all PLS variants are equivalent during the computation of the first dimension.

1.2 SVD decomposition and PLS-SVD

We recall the SVD decomposition and the principle of the PLS-SVD approach, that will be useful for understanding our sparse PLS approach.

1.2.1 Singular Value Decomposition

Any real r-rank matrix M $(p \times q)$ can decomposed into three matrices U, Δ, V as follows:

$$M = U \Delta V^T$$

where $U(p \times r)$ and $V(q \times r)$ are orthonormal and $\Delta(r \times r)$ is a diagonal matrix whose diagonal elements δ_k $(k = 1 \dots r)$ are called the singular values. The singular values are equal to the square root of the eigenvalues of the matrices $M^T M$ and $M M^T$. One interesting property that will be used in our sparse PLS method is that the columns vectors of U and V, noted $(u_1, \dots, u_r)$ and $(v_1, \dots, v_r)$ (resp. called left and right singular vectors) correspond to the PLS loadings of X and Y if $M = X^T Y$.

1.2.2 PLS-SVD

In PLS-SVD, the SVD decomposition of $M = X^TY$ is performed only once, and for each dimension h, M is directly deflated by its rank-one approximation ($M_h = M_{h-1} - \delta_h u_h v_h'$). This computationally attractive approach may however lead to non mutually orthogonal latent variables, a property of PLS2 ($\xi_s'\xi_r = 0, r < s$) and PLS-mode A ($\xi_s'\xi_r = 0$ *and* $\omega_s'\omega_r = 0, r < s$).

1.3 Soft-thresholding penalization

Shen and Huang (2007) proposed a sparse PCA approach using the SVD decomposition of $X = U\Delta V^T$ by penalizing the PCA loading vector v_k. The optimization problem to solve is

$$\min_{u,v} ||X - uv'||_F^2 + P_\lambda(v) \tag{3}$$

where $||X - uv'||_F^2 = \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - u_i v_j)^2$ and $P_\lambda(v) = \sum_{j=1}^p p_\lambda(|v_j|)$ is a penalty function. Among different penalty functions that were proposed, we considered the soft-thresholding function.

Solving (3) is performed in an iterative way, as described below:

- Decompose $X = U\Delta V^T$, $X_0 = X$
- For h in 1..H:
  1. Set $v_{old} = \delta_h v_h^\star$, $u_{old} = u_h^\star$, where $v_h^\star$ and $v_h^\star$ are unit vectors
  2. Until convergence of u_{new} and v_{new}:
     (a) $v_{new} = g_\lambda(X_{h-1}^T u_{old})$
     (b) $u_{new} = X^T v_{new} / ||X_{h-1}^T v_{new}||$
     (c) $u_{old} = u_{new}$, $v_{old} = v_{new}$
  3. $v_{new} = v_{new} / ||v_{new}||$
  4. $X_h = X_{h-1} - \delta_h u_{new} v_{new}'$

where $g(y) = sign(y)(|y| - \lambda)_+$ is the soft-thresholding penalty function.

In our particular PLS case, we are interested in penalizing both loadings vectors u_k and v_k to perform variable selection in both data sets. Indeed, one interesting property of PLS is the direct interpretability of the loadings vectors as a measure of the relative importance of the variables in the model (Wold et al., 2004). Our optimization problem becomes:

$$\min_{u,v} ||M - uv'||_F^2 + g_{\lambda_1}(u) + g_{\lambda_2}(v) \tag{4}$$

which is solved iteratively by replacing X by M and the steps 2.**a.** and 2.**b.** by:

$$v_{new} = g_{\lambda_1}(M_{h-1}^T u_{old})$$

$$u_{new} = g_{\lambda_2}(M_{h-1}v_{old})$$

The sparse PLS algorithm is detailed in next section.

1.4 Sparse PLS

It is easy to understand that during the deflation step of the PLS-SVD, $M_h \neq X_h^T Y_h$. This is why we propose to compute separately X_h and Y_h, then to decompose at each step $\tilde{M}_h = X_h^T Y_h$ and finally, to extract the first pair of singular vectors. As Hoskuldsson (1988) explains, taking one pair of loadings (u_h, v_h) at a time will lead to a biggest reduction of the total variation in the X and Y-spaces. In our approach, the SVD decomposition will provide a useful tool for selecting variables from each of the two-blocks data. We now detail the sparse PLS algorithm (*sPLS*) based on the iterative PLS algorithm (see Tenenhaus 1998) and SVD computation of $\tilde{M}_h$ for each dimension.

1. $X_0 = X \qquad Y_0 = Y$
2. For h in 1..H:
   (a) Set $\tilde{M}_{h-1} = X_{h-1}^T Y_{h-1}$
   (b) Decompose $\tilde{M}_{h-1}$ and extract the first pair of singular vectors $u_{old} = u_h$ and $v_{old} = v_h$
   (c) Until convergence of u_{new} and v_{new}:
      i. $u_{new} = g_{\lambda_2}(\tilde{M}_{h-1}v_{old})$, norm u_{new}
      ii. $v_{new} = g_{\lambda_1}(\tilde{M}_{h-1}^T u_{old})$, norm v_{new}
      iii. $u_{old} = u_{new}$, $v_{old} = v_{new}$
   (d) $\xi_h = X_{h-1}u_{new}/u'_{new}u_{new}$
   $\omega_h = Y_{h-1}v_{new}/v'_{new}v_{new}$
   (e) $c_h = X_{h-1}^T\xi_h/\xi'_h\xi_h$
   $d_h = Y_{h-1}^T\xi_h/\xi'_h\omega_h$
   $e_h = Y_{h-1}^T\omega_h/\omega'_h\omega_h$
   (f) $X_h = X_{h-1} - \xi_h c'_h$
   (g) $Y_h = Y_{h-1} - \xi_h d'_h$

Note that in the case where there is no sparsity constraint ($\lambda_1 = \lambda_2 = 0$) we obtain the same results as in a classical PLS.

1.5 Missing data

When dealing with biological data, it is very common to be confronted to missing data. In order not too lose too much information, an interesting approach to substitute each missing data with a value can be the Non LInear Estimation by Iterative Partial Least

Squares (NIPALS, Wold 1966). This method has been at the origin of PLS and allows performing PCA with missing data on each block data set. Details of the algorithm can be found in Tenenhaus (1998). Several studies show that the convergence of NIPALS and its good estimation are limited by the number of missing values (20-30%), see for example Dray et al. (2003). NIPALS is now implemented in the `ade4` package.

1.6 Tuning criteria and evaluation

1.6.1 Soft-thresholding penalization

The two penalization parameters λ_1^h and λ_2^h can be simultaneously chosen by computing the error prediction ("*RMSEP*" see section 1.6.3) with k-fold cross validation or leave-one-out cross validation, and this for each given dimension h. In practice however, when analyzing biological data, our experience showed that an optimal tuning of the penalization parameters by optimizing the predictive ability of the model, does not necessarily satisfy the biologists needs. Indeed, in biological data sets, many omics data are still unknown (*e.g* associated functions, annotations) and too small variable selections might not allow for the biologists to correctly assess the results. This is why they may prefer instead to choose the number of non zero components in each loading vector u_h, v_h or in both, for each dimension h. This option was proposed in Zou and Hastie (2005) in their `R` package `elasticnet` for their sparse PCA.

1.6.2 Choice of PLS dimension

Marginal contribution of the latent variable ξ_h. In the case of a regression context, Tenenhaus (1998) proposed to compute a criteria called Q_h^2 that measures the marginal contribution of ξ_h to the predictive power of the PLS model, by performing cross validation computations. Here, as the number of samples n is usually small, we propose to use leave-one-out cross validation (loo-cv). Q_h^2 is computed for all Y variables and is defined as

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{q} PRESS_{kh}}{\sum_{k=1}^{q} RSS_{k(h-1)}},$$

where $PRESS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{h(-i)}^k)^2$ is the PRediction Error Sum of Squares and $RSS_h^k = \sum_{i=1}^{n}(y_i^k - \hat{y}_{hi}^k)^2$ is the Residual Sum of Squares for the variable k and the PLS dimension h.

We define the estimated matrix of regression coefficients $\hat{B}$ of B, using the same notation as in equation (2): $\hat{B} = U^* D^T$ where $U^* = U(C^T U)^{-1}$ (see De Jong and Ter Braak 1994; Tenenhaus 1998) and where the column vectors of U are the loading vectors $(u_1, \ldots, u_h)$, $h = 1 \ldots H$. For any i sample, we can predict $\hat{y}_{hi}^k = x_{hi}\hat{B}_{h(-i)}^k$.

This criteria is the one adopted in the *SIMCA-P* software (developed by S. Wold and Umetri 1996). The rule to decide if ξ_h contributes significantly to the prediction is if

$$Q_h^2 \geq (1 - 0.95^2) = 0.0975$$

However, the choice of the PLS dimension still remains an open question that has been mentioned by several authors (see Mevik and Wehrens 2007; Boulesteix 2004). In our particular biological context, we can show that graphical representations of the samples facilitate this choice as the plots of (ξ_h, ξ_{h+1}) and (ω_h, ω_{h+1}) do not have a biological meaning if h is too large. In fact, our results (see below) show that all relevant information in terms of identification of the measured biological effects can be extracted from 3 dimensions.

1.6.3 Evaluation

RMSEP For a regression context, Mevik and Wehrens (2007); Boulesteix (2004) in the `R pls` and `plsgenomics` packages proposed to compute the Root Mean Squared Error Prediction criterion (RMSEP) with cross validation in order to choose the H parameter. As we already suggested to use the Q_h^2 criterion for this issue, we propose instead to use the RMSEP criterion as a way of evaluating the predictive power for each Y variable between the original non-penalized PLS and the sPLS in the next section.

Note that the Q_h^2 criteria is closely related to RMSEP ($= PRESS_{kh}$) and gives a more general insight of the PLS, whereas the RMSEP requires to be computed for each variable k in Y.

2 Validation studies

The evaluation of any statistical approach is usually performed with simulated data sets. In the context of biological data, however, simulation is a difficult exercise as one has to take into account technical effects that are not even easily identifiable on the real data sets. We first propose to simulate as realistically as possible two-block data sets in a regression framework, to answer the questions : does the sparse PLS select relevant variables ? Does the variable selection performed simultaneously on both data sets improve the predictive ability of the model, compared to the PLS that includes all variables in the model ? Once these questions are answered, we propose on the next step to show that our approach is applicable on biological data sets with various complexities, and that it may give potentially relevant results from a statistical point of view compared to PLS. Finally, in the next section, we provide a detailed biological interpretation for one of the data set, and show that sparse PLS answers the biological question compared to the PLS.

2.1 Simulation study

2.1.1 Simulation design

As proposed by Chun and Keles (2007), this simulation is designed to compare the prediction performance of the PLS and sPLS in the case where the relevant variables are not governed by a latent variable model. In this setting, we also added two cross conditions to complexify this setting. We set $p = 5000$ genes, $q = 50$ response variables and $n = 40$ samples, all with base error model being Gaussian with unit variance. We defined the mean vectors μ_1 and μ_2 as follows and divided the samples into consecutive blocks of 10, denoted by the sets (a, b, c, d), where

$$\mu_{1i} = \begin{cases} -2 & \text{if } i \in a \cup b \\ +2 & \text{otherwise.} \end{cases}$$

$$\mu_{2i} = \begin{cases} -1.5 & \text{if } i \in a \cup c \\ +1.5 & \text{otherwise.} \end{cases}$$

For the first 20 genes, we generated 20 columns of X from a multivariate normal with an AR(1) covariance matrix with auto correlation $\rho = 0.9$. These genes will get a strong Y response, but should not be of interest in the model. The next 40 genes have the mean structure μ_1 or μ_2:

$$x_{ij} = \mu_{1i} + \epsilon_{ij}, \quad j = 21 \dots 40, \quad i = 1 \dots n.$$

$$x_{ij} = \mu_{2i} + \epsilon_{ij}, \quad j = 41 \dots 60, \quad i = 1 \dots n.$$

The next genes have the mean structure U_m and are generated by $X_j = U_m + \epsilon_j$, $m = 1 \dots 4$,

$$\begin{aligned} U_1 = &\quad -1.5 + 1.5\mathbb{1}_{u_{ij} \leq 0.4}, \quad 1 \leq i \leq n, \quad 61 \leq j \leq 80, \\ U_2 = &\quad +1.5 - 1.5\mathbb{1}_{u_{ij} \leq 0.7}, \quad 1 \leq i \leq n, \quad 81 \leq j \leq 100, \\ U_3 = &\quad -2 + 2\mathbb{1}_{u_{ij} \leq 0.3}, \quad 1 \leq i \leq n, \quad 101 \leq j \leq 120, \\ U_4 = &\quad +2 - 2\mathbb{1}_{u_{ij} \leq 0.3}, \quad 1 \leq i \leq n, \quad 121 \leq j \leq 140, \end{aligned}$$

where $u_{ij} \sim \mathcal{U}(0, 1)$ and ϵ_j are i.i.d random vectors from $\mathcal{N}(0, \mathbb{1}_n)$. In all cases, $\epsilon_{ij} \sim \mathcal{N}(0, 1)$, which is also how the remaining 4860 genes are defined.

The response variables Y_{ik} follow $Y_k = X\beta_1 + e_k$, $k = 1 \dots 10$, with

$$\beta_{1j} = \begin{cases} 10 & \text{if } 1 \leq j \leq 20 \\ 8 & \text{if } 21 \leq j \leq 40 \\ 4 & \text{if } 21 \leq j \leq p, \end{cases}$$

and $Y_k = X\beta_2 + e_k$, $k = 11 \dots 20$ with

$$\beta_{2j} = \begin{cases} 10 & \text{if } 1 \leq j \leq 20 \\ 4 & \text{if } 21 \leq j \leq 40 \\ 8 & \text{if } 21 \leq j \leq p \end{cases}$$

Table 1: Averaged RMSEP (standard error) for each PLS dimension for 50 simulated data sets.

	PLS	sparse PLS
dim 1	0.930 (0.009)	0.715 (0.030)
dim 2	0.927 (0.009)	0.581 (0.019)
dim 3	0.926 (0.009)	0.580 (0.019)

and $Y_k \sim e_k$ for $k = 21 \dots 50$ with $e_k \sim \mathcal{N}(0, \mathbb{1}_n)$.

In this simulation setting, the tested methods should highlight the genes X_j, $j = 11 \dots 40$ and the response variables Y_k, $k = 1 \dots 30$, which are related either to a μ_1 or μ_2 effect.

2.1.2 Prediction performance

X and Y are simulated 50 times and we use 10-fold cross validation on each data set. For the sparse PLS, we arbitrarily chose to select 50 genes and 30 response variables for each dimension h, $h = 1 \dots 3$. For PLS, no penalization is applied, so that all Y variables are modelled with respect to the whole X data set for each simulation run.

The RMSEP for each response variable, each test set and each dimension is computed and averaged in Table 1. These first results show that sparse PLS improves the predictive ability of the model. After dimension $H = 2$, neither sPLS nor PLS get a significantly decreasing averaged RMSEP. This is in agreement with our simulation design, in which only two latent effects, the μ_1 and μ_2 effects, are included. The next section show that these effects are indeed highlighted by PLS and sPLS in the first 2 dimensions.

2.1.3 Variable selection

In this part, we compare the loading vectors (u_1, u_2, u_3) and (v_1, v_2, v_3) in the PLS and the sPLS in one simulation run (results were similar for the other runs). Figure 1 shows that both PLS and sparse PLS highlight the "good" genes, but with no clear distinction between the group of genes with a μ_1 or μ_2 effect for the PLS in dimension 1 or 2. On the contrary, the sparse PLS clearly selects the μ_1 effect genes on dimension 2 with heavy weights. This may be useful for the biologists who want to clearly separate the genes related to each effect on a different dimension. For both methods, the dimension 3 does not seem to be informative. The same conclusion can be drawn for the Y variables.

If an artificial two step selection procedure is performed in PLS, first by ordering the absolute values of the loadings and then selecting a chosen number of variables, to select 50(30) genes (response variables) for the first three dimensions, the two selections in PLS and sPLS are roughly the same (identical for dimension 1, up to 5 different selected variables in dimension 2 and 3). This shows that sPLS simply seems to shrink the PLS

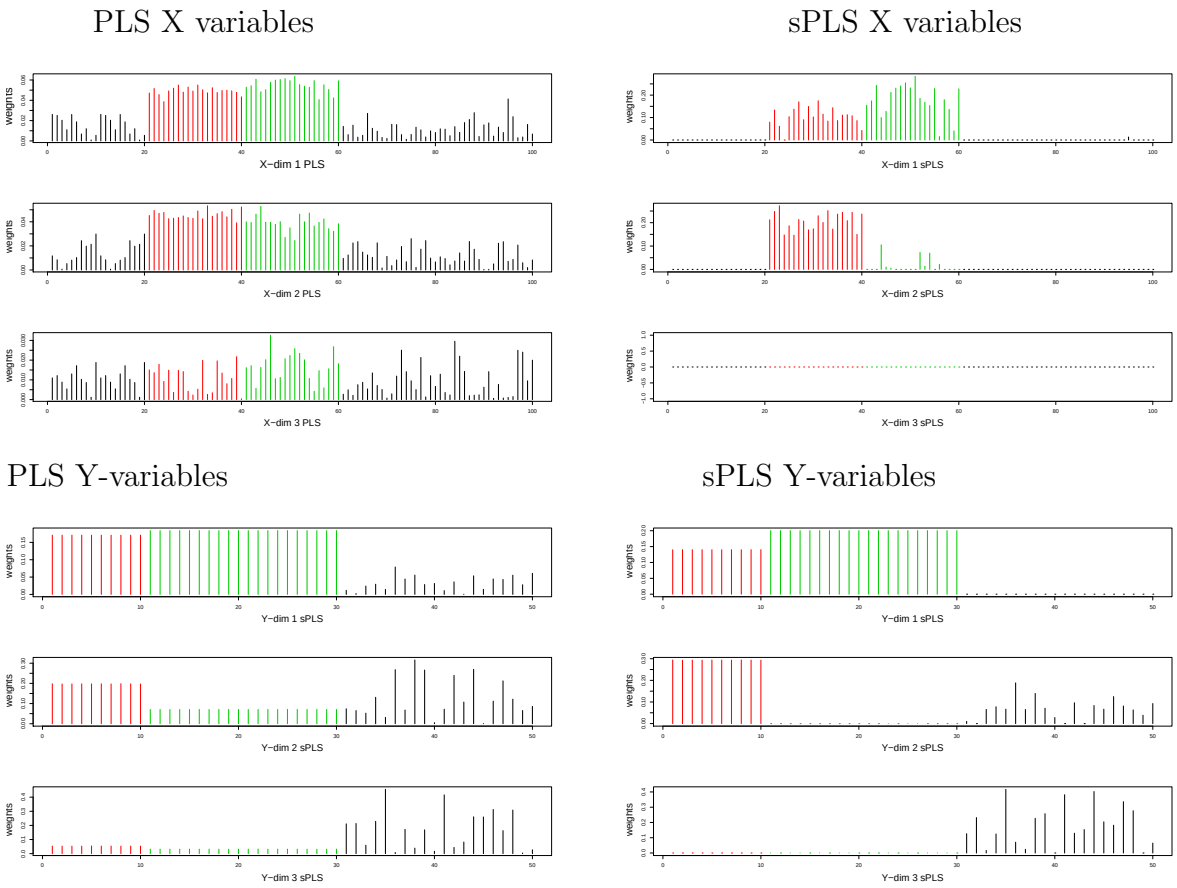

Figure 1: Absolute variable weights in the loading vectors of PLS (left) or sparse PLS (right) for the first 100 X variables (top) and the Y variables (bottom). The whole X variables weights can be found in supplementary material. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.

Table 2: Description of the data sets.

	Liver Toxicity	Arabidopsis	Wine Yeast
# samples n	64	18	43
X	gene expression	transcript	transcript
p	3116	22 810	3381
Missing values	2	0	0
Y	clinic variables	metabolite	metabolite
q	10	137	22
Missing values	0	22	0

loading coefficients in this simple controlled setting. However, on real data sets (see below), the difference between the two methods is genuine in terms of variable selection.

2.2 Case studies

2.2.1 Data sets

Liver Toxicity study In the liver toxicity study (Heinloth et al., 2004), 4 male rats of the inbred strain Fisher 344 were exposed to non-toxic (50 or 150 mg/kg), moderately toxic (1500 mg/kg) or severely toxic (2000 mg/kg) dose of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from the liver was extracted. Ten clinical chemistry measurements variables containing markers for liver injury are available for each object and numerically measure the serum enzymes level. The expression data are arranged in a matrix X of $n = 64$ objects and $p = 3116$ expression levels after normalization and pre-processing (Bushel et al., 2007). There are 2 missing values in the gene expression matrix.

In the original descriptive study, the authors claim that the clinical variables might not help detecting the paracetamol toxicity in the liver, but that the gene expression could be an alternative solution. However, in a PLS framework, we can be tempted to predict the clinical parameters (Y) by the gene expression matrix (X), as performed in Gidskehaug et al. (2007).

Arabidopsis data The responses of 22810 transcript levels and 137 metabolites and enzymes (including 67 unidentified metabolites) during the diurnal cycle (6) and an extended dark treatment (6) in WT Arabidopsis, and during the diurnal cycle (6) in starch less pgm mutants, is studied (Gibon et al., 2006). The aim is to detect the change of enzyme activities by integrating the changes in transcript levels and detect the correlation between the different time points and the 3 genotypes.

According to this previous study, metabolites and enzymes are regulated by gene

expressions rather than vice versa. We hence assigned to the X matrix the transcript levels and to the Y matrix the metabolites. The Y data set contained 22 missing values. This data set is characterized by a very small number of samples (18).

Wine Yeast data set *Saccharomyces cerevisiae* is an important component of the wine fermentation process and determines various attributes of the final product. One such attribute that is important from an industrial wine-making perspective is the production of volatile aroma compounds such as higher alcohols and their corresponding esters (Nykanen and Nykanen, 1977; Dickinson et al., 2003). The pathways for the production of these compounds are not clearly delineated and much remains unknown regarding the roles and kinetics of specific enzymes. In addition, most of the key reactions in the various pathways are reversible and the enzymes involved are fairly promiscuous regarding substrate specificity (Bely et al., 1990; Ribéreau-Gayon et al., 2000). In fact, different yeast strains produce wines with highly divergent aroma profiles. The underlying genetic and regulatory mechanisms responsible for these differences are largely unknown due to the complex network structure of aroma-producing reactions. As such an unbiased, holistic systems biology approach is a powerful tool to mine and interpret gene expression data in the context of aroma compound production.

In this study, five different industrial wine yeast strains (VIN13, EC1118, BM45, 285, DV10) were used in fermentation with synthetic must, in duplicate or triplicate (biological repeats). Samples were taken for microarray analysis at three key time points during fermentation, namely Day2 (exponential growth phase), Day5 (early stationary phase) and Day14 (later stationary phase). Exometabolites (aroma compounds) were also analysed at the same time by GC-FID.

Microarray analysis was carried out using the Affymetrix platform, and all normalizations and processing was performed according to standard Affymetrix procedures. To compensate for the bias towards cell-cycle related genes in the transcriptomic data set, the data was pre-processed to remove genes that are exclusively involved in cell cycle, cell fate, protein bio synthesis and ribosome bio genesis, leaving a set of 3391 genes for a regression framework analysis, with no missing data, and $n = 43$ samples.

2.2.2 Comparisons with PLS

Comparisons with PLS will be performed in terms of criteria defined in section 1.6: Q^2_h, predictive power assessment of the model as well as insight into the variable selection in terms of stability. As the main objective of this paper is to show the feasibility of the sparse approach, the three data sets will be used as illustrative examples to compare PLS and sPLS. In this regression framework, some of the data sets are characterized by a very small q (Liver Toxicity: $q = 10$, Wine Yeast $q = 22$). In these cases, we did not judge relevant to perform selection on these Y variables, and hence $\lambda^h_2 = 0$. In the other data set Arabidopsis, the selection was simultaneously performed on the X and Y data sets, as initially proposed by our approach.

Each input matrix was centered to column mean zero, and scaled to unit variance

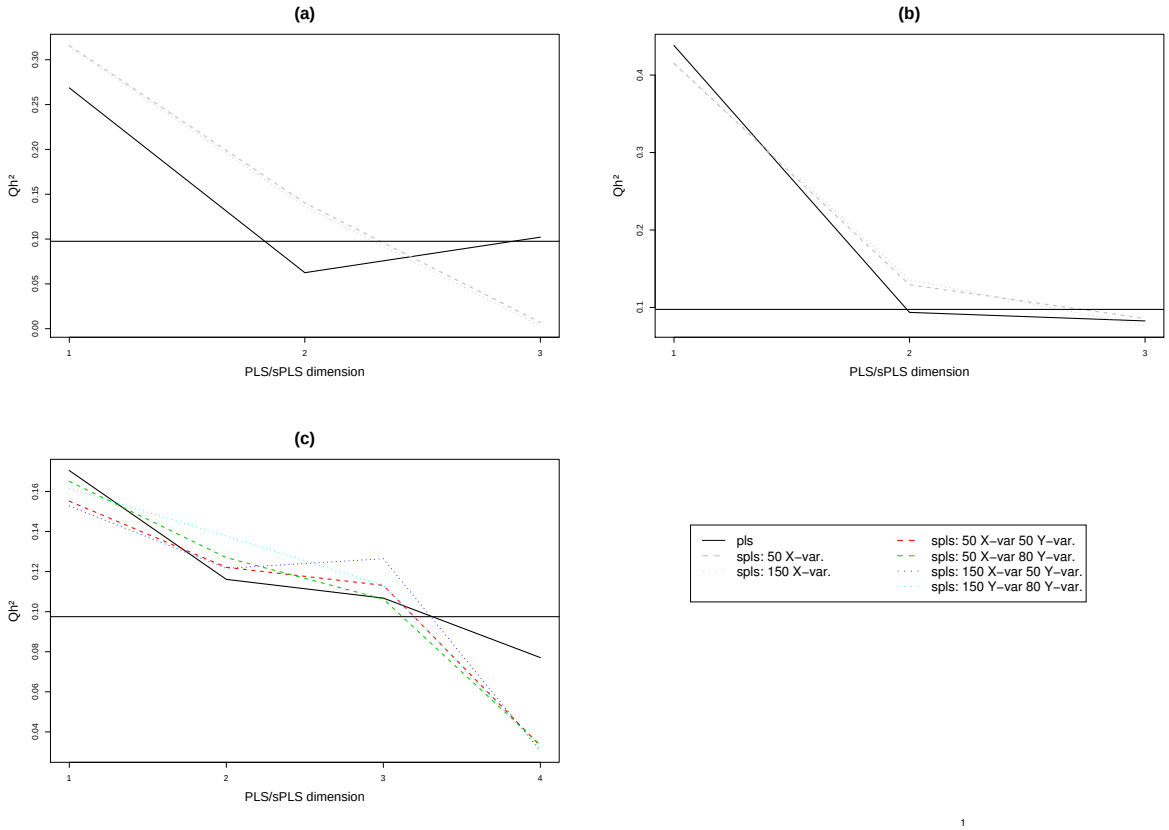

Figure 2: Marginal contribution of the latent variable ξ_h for each component in PLS and sPLS for different sparsity degrees for Liver Toxicity Study **(a)**, Wine Yeast **(b)** and Arabidopsis **(c)**. The horizontal black line indicates the threshold value in Q_h^2.

so as to avoid any dominance of one of the two-block data sets. Missing values were imputed with the NIPALS algorithm.

Q_h^2. We compare the Q_h^2 value with the PLS model with all variables in the model, and the proposed sparse PLS model with different sparsity degrees on each dimension : selection of 50 or 150 X variables on Liver Toxicity and Wine Yeast, 50 or 150 X variables coupled with the selection of 50 or 80 Y-variables in Arabidopsis. The choice of the selection size is arbitrarily chosen and loo-cv is applied for all data sets. The marginal contribution of ξ_h for each PLS/sPLS component is computed for each dimension. Figure 2 shows that the values of Q_h^2 behave differently, depending on the data set and on the PLS/sPLS approach.

In Liver Toxicity and Wine Yeast **(a) (b)**, PLS needs one less component than sPLS : 1 (2) PLS dimensions for Liver Toxicity (Wine Yeast). As already observed in section 2.1.3, sPLS would need one more dimension to fully separate the different biological effects and select the X and Y variables according to each of these effects.

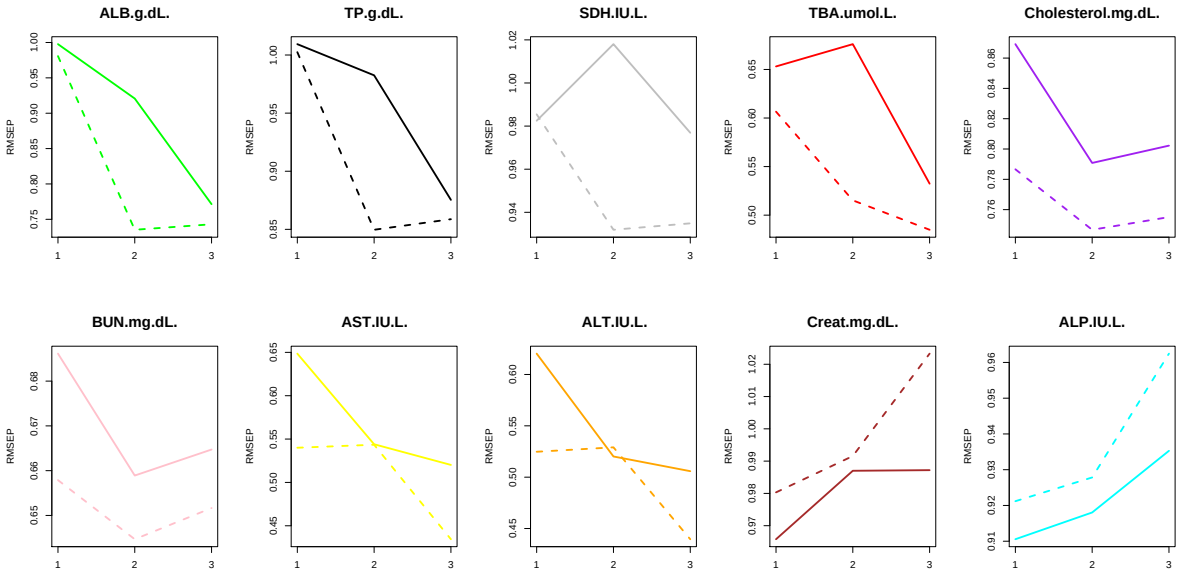

Figure 3: Liver Toxicity study: RMSEP for each clinical variable with PLS (plain line) and sPLS (dashed). Clinical variables are ranked according to their loadings in dimension 2.

In Liver Toxicity, Q^2_3 increases and becomes superior to the threshold value 0.0975. On the other hand, the Q^2_h values in any sPLS steadily decreases with h.

In Arabidopsis **(c)** that is characterized by many X variables, and where a simultaneous variable selection is performed on the Y data set, the Q^2_h values differ depending on the number of variables that are selected on both data sets. However, for both methods and all sparsity degrees, the choice of $H = 3$ seems sufficient.

Predictive ability. Figure 3 compares the RMSEP for each clinical variable in the Liver Toxicity study with PLS (no selection) and sPLS (here, selection of 150 genes). As observed in section 2.1.2, these graphics show that except for 2 clinical variables, sPLS clearly outperforms PLS. Removing some of the noisy variables in the X data set helps for a better prediction of most of the Y variables. In this figure, the clinical variables are ranked according to the absolute value of their loadings in v_2. Hence the Y-loadings have a meaning in terms of variable importance measure, as the less better explained variables *creat.mg.dL* and *ALP.IU.L* get the lowest ranks. A thorough biological interpretation would be needed here to verify if these clinical variables are relevant in the biological study.

If the clinical variables were ranked according to the next loading v_3, then, although the graphics would be unchanged, *creat.mg.dL* and *ALP.IU.L* would get a a higher rank (resp. rank 1 and 8). This result comforts the choice of $H = 2$ for Liver Toxicity with sPLS. Similar conclusions can be drawn on the other data sets that includes more Y variables.

Table 3: Stability : ratio of the true positive variables selected in original data set and bootstrap data sets over the length of each selection (100).

	Liver toxicity		Arabidopsis				Wine Yeast	
	PLS	sPLS	PLS		sPLS		PLS	sPLS
			X	Y	X	Y		
dim 1	0.735	0.739	0.332	0.895	0.377	0.893	0.596	0.598
dim 2	0.457	0.603	0.221	0.834	0.365	0.838	0.622	0.559
dim 3	0.354	0.279	0.101	0.77	0.156	0.78	0.52	0.463

Table 4: Number of variables commonly selected in PLS (two step selection procedure) and in sPLS when selecting 100 variables.

	Liver toxicity	Arabidopsis		Wine Yeast
	X	X	Y	X
dim 1	97	56	90	91
dim 2	56	45	82	73
dim 3	19	72	80	74

Variable selection.

Stability. On B bootstrap samples, $B = 10$, we compare the 100 X variables and 100 Y variables (in the case of Arabidopsis) that were selected either with PLS (two step selection procedure) or sPLS with respect to the same number of variables selected on the original data sets. The results are summarized in Table 3 and show that except for Wine Yeast in dimension 2 and 3, the sparse PLS approach seems more stable than PLS. It is not surprising to find an increased stability when the total number of variables (p and q) is rather small.

Comparison with PLS. Table 4 highlights the actual differences between a selection performed either with PLS (in two steps) or with sPLS for the same number of variables (100 for each data set, when applicable). As expected, both selections should be similar in dimension 1, but differ greatly for the other dimensions. In particular, the selections performed in the X Arabidopsis data set differ from the very first dimension. This is due to the extremely large number of X variables ($p = 22810$), where many of the transcripts get similar weights in PLS.

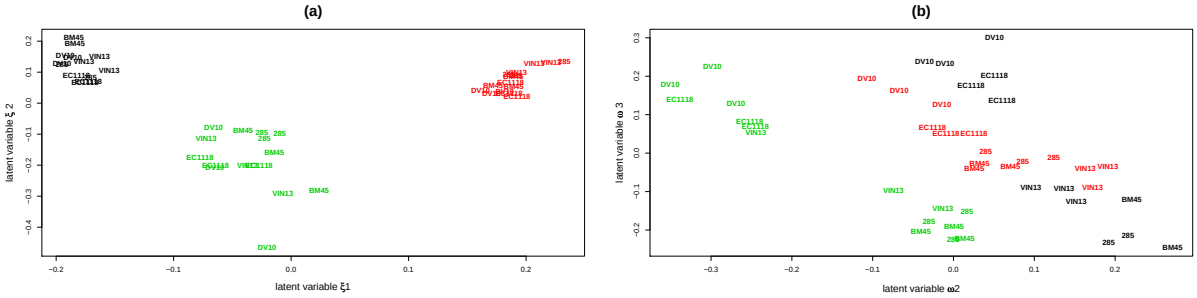

Figure 4: Wine Yeast data : graphical representation of the samples for the latent vectors (ξ_1, ξ_2) **(a)** and (ω_2, ω_3) **(b)**. Colors red, green and black stand for fermentation day 2, 5 and 14.

2.2.3 Property of the loading vectors.

When applying sparse methods, the loadings may lose their property of orthogonality and uncorrelation, as it was observed with sparse PCA (Trendafilov and Jolliffe, 2006; Shen and Huang, 2007). This is not the case with sPLS. In the original PLS, no constraint is set to have $\omega_r'\omega_s = 0$, $r < s$. Hence, latent variables $(\omega_1, \ldots, \omega_H)$ from the Y data set are not orthogonal in PLS or sPLS. To remedy to this in terms of graphical representation of the samples, we propose to re project $(\omega_1, \ldots, \omega_H)$ in an orthogonal basis. For the latent variables ξ, however, we always observed that $\xi_r'\xi_s = 0$ and no re projection is needed for these latent variables.

3 Analysis of the Wine Yeast data set and biological interpretation

We first give some elements of discussion regarding the graphical representation of the latent variables (samples), which facilitate the biological interpretation. These preliminary remarks will explain some of the results obtained when we compared the genes selected with PLS (two-step procedure) to the genes selected in the one step procedure with sPLS. Finally we show that the sPLS selection gives meaningful insight into the biological study.

As required by the biologists who performed this experiment, 200 genes were selected for each dimension.

3.1 Biological samples

Figure 4 highlights several facts that can actually be explained by the biological experiment. The plots of (ξ_1, ξ_2) (top) and (ω_1, ω_2) gave similar representation (not shown). The first component separated samples into time-specific clusters. This is to be expected as the particular stage of fermentation is the major source of genetic variation

Table 5: Comparison of genes selected with PLS (two step procedure) vs. sPLS.

	PLS	sPLS
dim 1	-genes related to general central carbon metabolism -inclusion of many dubious/suspect ORFs	-GDH1: key regulator of cellular redox balance (direct influence on the main aroma producing reactions
dim 2	-identifies 'rate-limiting' enzymes in aroma metabolism	-improved coverage of transcriptional pathways
dim 3	-identifies most important alcohol and aldehydes dehyrogenase genes	-IDH1: key enzyme controlling flux distribution between aroma producing pathways and TCA cycle -NDE1: provides energy intermediates for dehydrogenase reactions

and the main determinant of aroma compound levels. The next most significant source of biological variation is the identity of the yeast strain. This was corroborated by the second and third components, where the samples clustered together in biological repeats of the same strain. Strains that are known to be more similar in terms of their fermentative performance also clustered closely within time (*i.e* EC1118 and DV10, and BM45 and 285). The VIN13 strain (which is least similar to any of the other strains in this study) showed an intermediate distribution between the latent variable axes.

3.2 Selected variables

Comparisons with PLS Table 5 presents the similarities and main differences observed between the genes selected either with PLS or sPLS in regression mode. We adopted a two-step procedure to select genes with the original PLS approach by ordering the absolute values of the loadings u_h for each dimension ($H = 3$) and selecting the same number of top genes as in sPLS.

The striking result that we observed was the differences in the genes selections, especially in dimension 2 and 3. Overall, these dimensions were found to be more enriched for genes with proved or hypothesized roles in aroma compound production (based on pathway analysis and functional categorisation) for the sPLS rather than PLS.

Genes selected with sPLS. Figure 5 depicts the 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. From the figure it is clear that the sPLS outputs provided good coverage of key reactions and major branches of the aroma production pathways (for the areas of metabolism with known reactions and enzymes). The first component identified mostly genes that are involved in reactions that produce the key substrates for starting points of the pathways of amino acid degradation and higher alcohol production. Amino acid metabolism is also a growth stage-specific factor (linked to fermentative stage), which is supported by the observations discussed in section 3.1. Most of the crucial

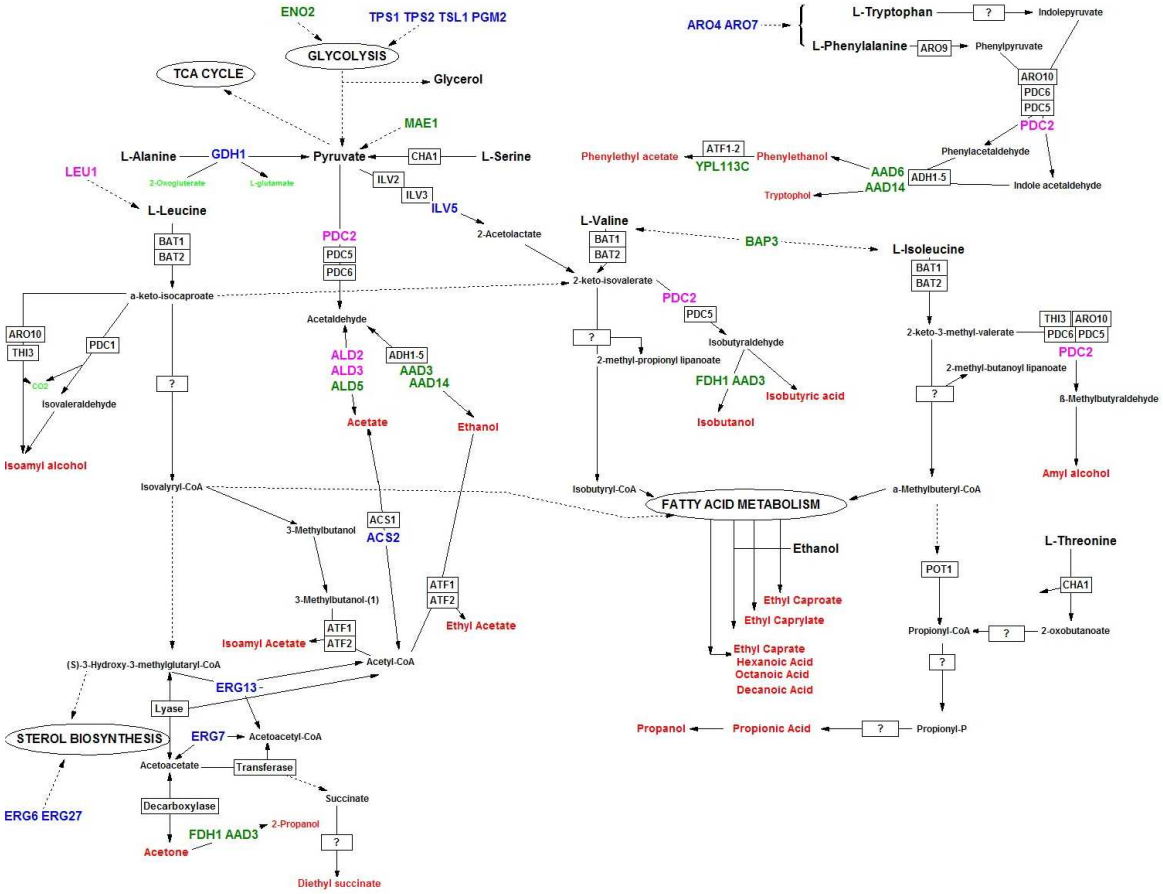

Figure 5: Graphical representation of 'known' or hypothesised reactions and enzyme activities involved in the reaction network of higher alcohol and ester production. Indirect interactions (*i.e* missing intermediates) are indicated by dashed lines and standard reactions are indicated by solid lines. Aroma compounds (red) and other metabolic intermediates (black) are positioned at the arrow apices. Unknown enzyme activities are represented by a question mark (**?**). Gene names coding for the relevant enzymes are represented in black box format, except for those genes that were identified in the first (blue), second (purple) and third (green) components of the sPLS.

'rate limiting' enzymes (PDC2, ALD2, ALD3, LEU1) were identified by the second component. In total, the highest number of relevant genes were identified by the third component. Genes in this component were also interesting from the perspective that they only have putative (but unconfirmed) roles to play in the various pathways where they are indicated in the figure. Associations between genes with putative functional designations (based on homology or active site configuration) and aroma compounds in the lesser annotated branches of aroma compound production provide opportunities for directed research and the formulation of novel hypothesis in these areas.

Further analysis to be done. An attractive way of representing variables is to compute the correlation between the original data set (X and Y) and the latent variables $(\xi_1, \ldots \xi_H)$ and $(\omega_1, \ldots \omega_H)$, as it is done with PCA or CCA. These graphical representations, where the selected variables are projected on a correlation circle, will allow to identify known and unknown relationships between the X variables, the Y variables, and more importantly between both types of omics data. Of course these relationship will then need to be biologically assessed with further experiments, and will constitute a next step of our proposed analysis.

4 Conclusion

We have introduced a general computational methodology that modifies PLS, a well known approach that has been proved to be extremely efficient in many data where $n << p + q$, in a sparse version including variable selection to be more useful to the biologists. Validation of the sparse PLS approach has been performed both on simulated but also on real data sets and compared with PLS. The simulation study showed that sPLS selected the relevant variables from both data sets, that were governed by the known latent effects. The application to real data sets showed that this built-in variable selection procedure improved the predictive ability of the model, differed from PLS from dimension 2 and seemed more stable. Compared to PLS, sPLS seemed to highlight each latent biological effect on a different dimension and accordingly select the variables governed by each effect. This result will help biologists identifying relevant variables linked to each biological condition.

Our proposed algorithm is fast to compute. Like any sparse multivariate method, sPLS requires the addition of penalization parameters. The tuning of these two parameters can simply be performed by choosing the variable selection size, a useful option for the biologists. The gain by penalizing, and hence selecting variables, is proved on a typical biological study aiming at integrating gene expressions and metabolites in Wine Yeast. We provide a thorough biological interpretation and show that the sPLS results are extremely meaningful for the biologist, compared to a PLS selection. This preliminary work undoubtedly brought more insight into the biological study and will suggest further experiments to be performed.

Integrating omics data is an issue that may soon be commonly encountered in most

high throughput biological studies. Hence we believe that our sparse PLS provides an extremely useful tool for the biologist in need of integrating two-block data sets and easily interpreting the resulting variable selections.

Remark. Another variant in our sparse PLS approach can be considered in step (g) of the proposed algorithm in section 1.4, by deflating the Y matrix in a symmetric manner: $Y_h = Y_{h-1} - \omega_h e_h'$. In this case, we are in a canonical framework and the aim is to model a *reciprocal relationship* between the two sets of variables. The lack of statistical criteria in this setting (as we are not in a predictive context) would require a thorough biological validation of the approach, rather than a statistical validation, and will constitute the next step of our research work.

Availability The code sources of sparse PLS (in R[1]) can be available upon request to the corresponding author. An R package is currently being implemented.

Acknowledgement

We are very grateful to Yves Gibon (Max Planck Institute of Molecular Plant Physiology) who kindly provided the full Arabidopsis data set.

References

Bely, M., Sablayrolles, J., and Barre, P. (1990). Description of Alcoholic Fermentation Kinetics: Its Variability and Significance. American Journal of Enology and Viticulture, 41(4):319–324.

Boulesteix, A. (2004). PLS Dimension Reduction for Classification with Microarray Data. Statistical Applications in Genetics and Molecular Biology, 3(1):1075.

Boulesteix, A. and Strimmer, K. (2005). Predicting transcription factor activities from combined analysis of microarray and chip data: a partial least squares approach. Theor Biol Med Model, 2(23).

Burnham, A., Viveros, R., and Macgregor, J. (1996). Frameworks for latent variable multivariate regression. Journal of chemometrics, 10(1):31–45.

Bushel, P., Wolfinger, R. D., and Gibson, G. (2007). Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. BMC Systems Biology, 1(15).

Bylesjö, M., Eriksson, D., Kusano, M., Moritz, T., and Trygg, J. (2007). Data integration in plant biology: the o2pls method for combined modeling of transcript and metabolite data. The Plant Journal, 52:1181–1191.

Chun, H. and Keles, S. (2007). Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis. Technical report, Department of Statistics, University of Wisconsin, Madison, USA.

Culhane, A., Perriere, G., and Higgins, D. (2003). Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4(1):59.

Datta, S. (2001). Exploring relationships in gene expressions: A partial least squares approach. Gene Expr, 9(6):249–255.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18:251–263.

[1]The Comprehensive R Archive Network, http://cran.r-project.org/

De Jong, S. and Ter Braak, C. (1994). Comments on the PLS kernel algorithm. Journal of chemometrics, 8(2):169–174.

Dickinson, J., Salgado, L., and Hewlins, M. (2003). The Catabolism of Amino Acids to Long Chain and Complex Alcohols in Saccharomyces cerevisiae. Journal of Biological Chemistry, 278(10):8028–8034.

Doledec, S. and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31(3):277–294.

Dray, S., Pettorelli, N., and Chessel, D. (2003). Multivariate Analysis of Incomplete Mapped Data. Transactions in GIS, 7(3):411–422.

Gibon, Y., Usadel, B., Blaesing, O., Kamlage, B., Hoehne, M., Trethewey, R., and Stitt, M. (2006). Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. Genome Biology, 7:R76.

Gidskehaug, L., Anderssen, E., Flatberg, A., and Alsberg, B. (2007). A framework for significance analysis of gene expression data using dimension reduction methods. BMC Bioinformatics, 8(1):346.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1):389–422.

Heinloth, A., Irwin, R., Boorman, G., Nettesheim, P., Fannin, R., Sieber, S., Snell, M., Tucker, C., Li, L., Travlos, G., et al. (2004). Gene Expression Profiling of Rat Livers Reveals Indicators of Potential Adverse Effects. Toxicological Sciences, 80(1):193–202.

Hoskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics, 2(3):211–228.

Lê Cao, K.-A., Gonçalves, O., Besse, P., and Gadat, S. (2007). Selection of biologically relevant genes with a wrapper stochastic algorithm. Statistical Applications in Genetics and Molecular Biology, 6(:Iss. 1):Article 1.

Lorber, A., Wangen, L., and Kowalski, B. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(19-31):13.

Mevik, B.-H. and Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in r. Journal of Statistical Software, 18(2).

Nykanen, L. and Nykanen, I. (1977). Production of esters by different yeast strains in sugar fermentations. J. Inst. Brew, 83:30–31.

Pihur, V., Datta, S., and Datta, S. (2008). Reconstruction of genetic association networks from microarray data: A partial least squares approach. Bioinformatics.

Ribéreau-Gayon, P., Dubourdieu, D., Donche, B., and Lonvaud, A. (2000). Biochemistry of alcoholic fermentation and metabolic pathways of wine yeasts in Handbook of Enology, volume 1. John Wiley and Sons.

Shen, H. and Huang, J. Z. (2007). Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis, to appear.

Tenenhaus, M. (1998). La régression PLS: théorie et pratique. Editions Technip.

Trendafilov, N. and Jolliffe, I. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. Computational Statistics and Data Analysis, 50(1):242–253.

Trygg, J. and Wold, S. (2003). O2-pls, a two- block (x-y) latent variable regression (lvr) method with an integral osc filter. Journal of Chemometrics, 17:53–64.

Umetri, A. (1996). SIMCA-P for windows, Graphical Software for Multivariate Process Modeling. Umea, Sweden.

Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology, 7(1):3.

Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.

Wold, H. (1966). Multivariate Analysis. Academic Press, New York, Wiley, krishnaiah, p.r. (ed.) edition.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The PLS method–partial least squares projections to latent structures–and its applications in industrial RDP (research, development, and production). Technical report, Umea University.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B, 67(2):301–320.

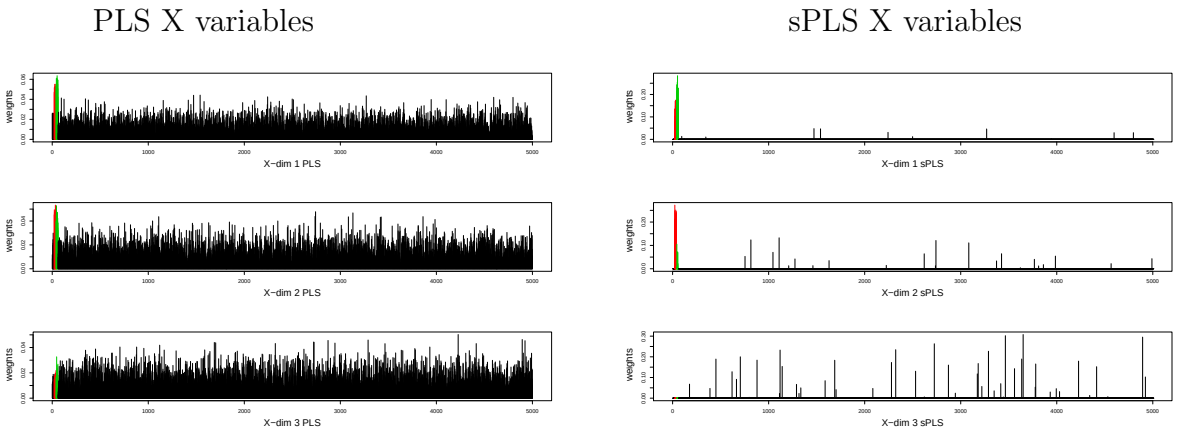

Figure 6: Supplementary figure: absolute variable weights in the loading vectors of PLS (left) or sparse PLS (right) for the 5000 X variables. Red (green) color stands for the variables related to the μ_1 (μ_2) effect.